P-55

e-Learning システムによる学習効果の測定 —Kullback-Leibler Divergence による確率分布推定の試み—

新潟医療福祉大学医療情報管理学科
近藤正紀

【背景・目的】

コンピューターを利用した学習において，応答時間（解答時間）から学習状態を把握する手法が提案されている．一方で，そもそも学習者の応答時間が従う確率分布については，ワイブル分布やガンマ分布などが提案されているもののはっきりしたことは未だ判っていない．

本稿では，応答時間分布として提案されている確率分布の多くが，一般化ガンマ分布の特別な場合であることを利用し，その母数を推定する方法を提案する．

【方法】

確率分布の一致度合いを測定する方法として，式 1 で与えられる Kullback-Leibler Divergence（以下，KL 距離）がよく用いられている．ただし，“距離”と呼ばれることが多いにもかかわらず，“距離の公理”を満たさないため，数学的な意味での距離ではない．

$$\mathrm{KL}(p_1 \mid p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \cdots\cdots (1)$$

一方，観測された頻度分布と理論分布の適合度検定に用いられる Pearson distance（PE 距離．式 2）も，最小二乗法と相性が良い等利点も多いが，距離の公理を満たさないのは KL 距離と同じである．

$$\mathrm{PE}(p_1 \parallel p_2) = \int p_2(x) \left(\frac{p_1(x)}{p_2(x)} - 1 \right)^2 dx \cdots\cdots (2)$$

距離の公理を満たし，KL 距離，PE 距離が内包する除算の非有界性の心配の無い距離も提案されているが，別の問題が生じる．過去にガンマ分布の母数を最尤推定法によって求めていることから，今回は最尤推定と相性の良い KL 距離を用いて，システムに残されている約 21 万件のログレコードから一般化ガンマ分布の母数の推定を試みた．

【結果】

通常のガンマ分布（式 3）の KL 距離は式 4 のようになる [1]．Γ(・)はガンマ関数，Ψ(・)はディガンマ関数である．

$$p(x \mid \alpha, \beta) = \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} e^{-\frac{x}{\beta}} \cdots\cdots (3)$$

$$\begin{aligned} \mathrm{KL}(p_1 \parallel p_2) = {} & (\alpha_1 - 1)\Psi(\alpha_1) - \log\beta_1 - \alpha_1 \cdots (4) \\ & -(\alpha_2 - 1)(\Psi(\alpha_1) + \log\beta_1) \\ & -\log\Gamma(\alpha_1) + \log\Gamma(\alpha_2) \\ & +\alpha_2 \log\beta_2 + \frac{\alpha_1\beta_1}{\beta_2} \end{aligned}$$

頻度分布を p_1，最尤推定法で得たガンマ分布を p_2 として，区分求積法で数値積分を行うことで頻度分布と推定したガンマ分布との KL 距離を計算できたことから，ガンマ分布の母数を KL 距離の最小化によって求めることにした．

KL 距離は距離の公理のうち非負性と非退化性のみを満たす．従って，KL 距離が 0 になる母数を求めることができれば頻度分布の理論分布が求められることになるが，実際には 0 になることは期待できないため，最小値（この場合は極小値）を求めることとなる．

過去の考察から推定値は $2 \le \alpha \le 4$，$5 \le \beta \le 7$ の範囲に収まるらしいと判っているが，差分 $\Delta\alpha = \Delta\beta = 0.1$ として各点での KL 距離を算出すると 861 個の KL 距離を算出してから最小値を求めることになる．精度を 10^{-6} とした場合は $\Delta\alpha = \Delta\beta = 10^{-7}$ であり，約 4×10^{12} 個の KL 距離を算出することになり現実的ではない．そこで初期値(α_0, β_0)から出発して，最急勾配法によって最小値（極小値）を求めることを試みた．

最急勾配法では，点(α_0, β_0)に対して 2 次元セルオートマトンで言うところの Moore 近傍各点における KL 距離を算出し，最小の KL 距離を与える点を推定値とし反復することで点列を収束させる．

ガンマ分布に対してこの方法を用いた場合は数分で求められることが判明した．

【考察】

一般化ガンマ分布には 3 母数のもの（式 5）と，更に一般化した 4 母数のもの（式 6）が提案されている．

$$p(x \mid \alpha, \beta, \gamma) = \frac{\gamma x^{\alpha\gamma-1}}{\beta^{\alpha\gamma}\Gamma(\alpha)} e^{-\left(\frac{x}{\beta}\right)^{\gamma}} \cdots\cdots (5)$$

$$p(x \mid \alpha, \beta, \gamma, \delta) = \frac{\gamma (x-\delta)^{\alpha\gamma-1}}{\beta^{\alpha\gamma}\Gamma(\alpha)} e^{-\left(\frac{x-\delta}{\beta}\right)^{\gamma}} \cdots\cdots (6)$$

3 母数の場合は KL 距離が得られている [2]が，4 母数については未知である．また，母数が 2 から 3 に増加する場合，現在のアルゴリズムで計算すると最悪のケースでは 100 万倍以上に計算量が増加する．4 母数の場合はその 2 乗となり，現実的な時間で求めることが困難になっている．

【結論】

本稿では e-Learning における応答時間分布を一般化ガンマ分布と仮定して母数を KL 距離の最小化によって推定することを目標とした．しかし，現有のアルゴリズム及び計算資源では算出が時間的に困難であることが示唆された．また，アルゴリズムの改良のためには，KL 距離の極小値近傍での挙動，及び，KL 距離が対称性と三角不等式を満たさないことの影響を考察する必要がある．

【文献】

1) Penny W D. KL-Divergence of Normal, Gammma, Dirichlet and Wishart densities. University College London. 2001.
2) Bauckhage C. Computing the Kullback-Leibler Divergence between two Generalized Gamma Distributions. University of Bonn. 2014.